National Library of Medicine

PubMed

| PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM |

Search PubMed ▾ for [   ]   Go   Clear

Limits   Preview/Index   History   Clipboard

Display   Abstract   ▾   Save   Text   Order   Add to Clipboard

Related Articles Books
LinkOut

FREE full text article at www.pnas.org

Highly specific protein sequence motifs for genome analysis.

Nevill-Manning CG, Wu TD, Brutlag DL

Department of Biochemistry, Stanford University, Stanford, CA 94305-5307, USA.

We present a method for discovering conserved sequence motifs from families of aligned protein sequences. The method has been implemented as a computer program called EMOTIF (http://motif.stanford.edu/emotif). Given an aligned set of protein sequences, EMOTIF generates a set of motifs with a wide range of specificities and sensitivities. EMOTIF also can generate motifs that describe possible subfamilies of a protein superfamily. A disjunction of such motifs often can represent the entire superfamily with high specificity and sensitivity. We have used EMOTIF to generate sets of motifs from all 7,000 protein alignments in the BLOCKS and PRINTS databases. The resulting database, called IDENTIFY (http://motif.stanford.edu/identify), contains more than 50,000 motifs. For each alignment, the database contains several motifs having a probability of matching a false positive that range from $10(-10)$ to $10(-5)$. Highly specific motifs are well suited for searching entire proteomes, while generating very few false predictions. IDENTIFY assigns biological functions to 25-30% of all proteins encoded by the Saccharomyces cerevisiae genome and by several bacterial genomes. In particular, IDENTIFY assigned functions to 172 of proteins of unknown function in the yeast genome.

PMID 9600885

Display   Abstract   ▾   Save   Text   Order   Add to Clipboard